U-Compare Apertium Part-of-Speech Tagging Workflow

1. BASIC INFORMATION

Tool name

UIMA/U-Compare Apertium Part-of-Speech Tagging Workflow

Overview and purpose of the tool

This is a workflow (processing pipeline) that is designed especially for use in the UIMA-based U-Compare workbench (Kano et al., 2009; Kano et al., 2011;see separate META-SHARE record).

The purpose of the workflow is to perform tokenisation, morphological analysis and part of speech tagging on plain text. The provided workflow can currently operate on a subset of the languages that are supported by the Apertium system, namely: English, Spanish, Calatan, Galician, Portuguese and Basque.

A short description of the algorithm

The workflow comprises the "Apertium Mopho" and "Apertium POS" UIMA[1] (Ferrucci et al., 2006) components, which are not part of U-Compare's core component library. These two components are modules of the Apertium machine translation system[2] (Armentano-Ollet et al., 2006).

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The workflow can be imported into the U-Compare platform[3] (see separate META-SHARE record). The only requirement to run U-Compare is for Java 6 to be installed.

Installation

In order to run the UIMA/U-Compare Apertium Part-of-Speech Tagging Workflow in U-Compare, it must be imported into U-Compare. Information about downloading and running U-Compare can be found in the documentation associated with the META-SHARE record for the U-Compare Workbench or the U-Compare website: http://nactem.ac.uk/ucompare/.

The workflow is in "ucz" format (a file format specific to U-Compare used for sharing workflows). The “Apertium_U-Compare_Tagging_Workflow.ucz” file should be downloaded and saved to disk. The following steps show how to import the workflow for use in U-Compare.

[1] http://uima.apache.org/

[2] http://www.apertium.org/

[3] http://nactem.ac.uk/ucompare/

1) From the “Workflow” menu in the U-Compare Workbench, choose the item “Import Workflow/Components”, as shown in in Figure 1.

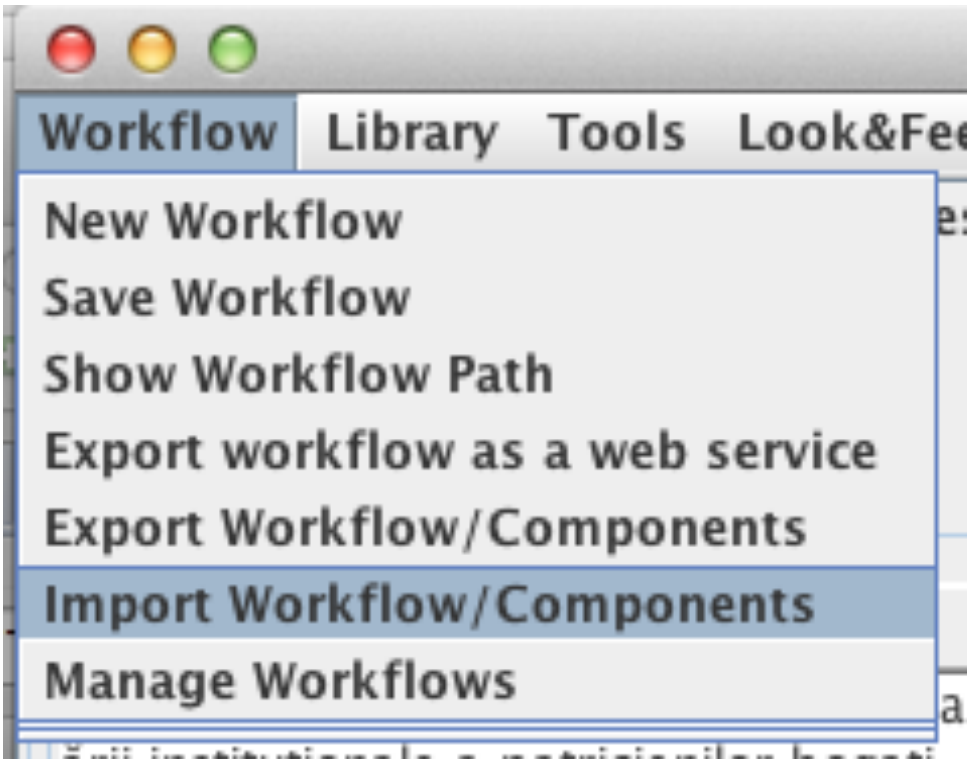
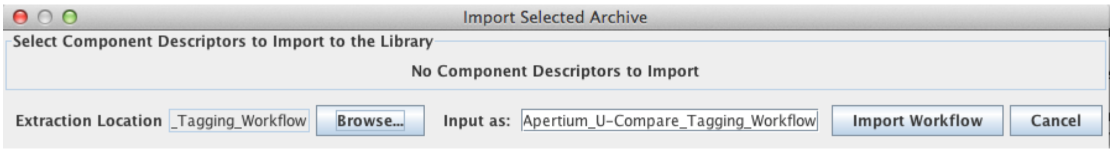


Figure 1: Workflow menu in U-Compare

2) A file chooser window will appear; select the location of the “Apertium_U-Compare_Tagging_Workflow.ucz” file, and click on “Open”.

3) A dialog box will appear allowing the workflow to be imported (Figure 2). A default location and name for the workflow will be specified; these can be changed if required. The “Import Workflow“ button should be clicked to proceed with the importing.



No Component Descriptors to Import

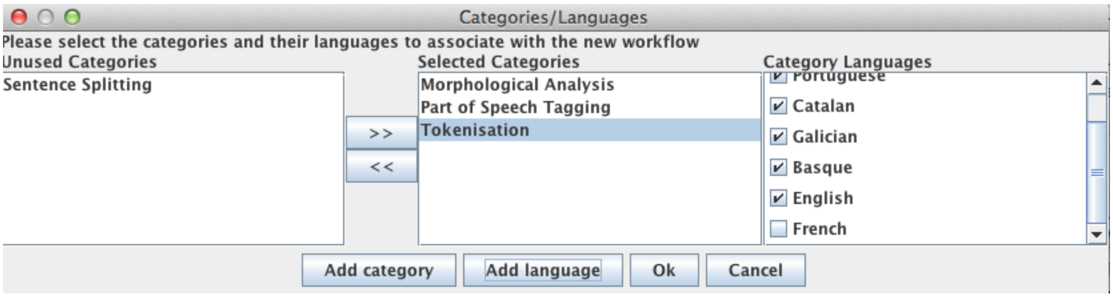
Extraction Location: _Tagging_Workflow

Input as: Apertium_U-Compare_Tagging_Workflow

Figure 2: Import workflow window

4) To help the user arrange workflows, U-Compare asks each workflow to be categorised according to its function and supported languages. Thus, a “Categories/Languages” (Figure 3) window will appear after the “Import Workflow” button has been clicked. A set of available workflow categories is shown under “Unused Categories”. The user can choose which of these categorie(s) the workflow falls within, by selecting them and clicking the right arrows button to transfer them to

the “Selected Categories” box. If the workflow performs a function that is not in the current list of categories, then the “Add Category” button can be clicked to add a new workflow category. In Figure 3, three workflow categories have been selected, i.e. “Morphological Analysis”, “Part-of-Speech tagging” and “Tokenisation”. Next, the supported languages for each workflow category should be chosen. When one of the “Selected Categories” of workflows is clicked upon, a list of languages is displayed in the “Category Languages” box. Any languages that the workflow supports should be checked. In case the workflow supports a language that is not in the list, the “Add Language” button can be clicked to add a new language to the list. After the appropriate categories/languages have been selected, the “OK” button should be clicked, which completes the import procedure.


Categories/Languages

Please select the categories and their languages to associate with the new workflow

Unused Categories: Sentence Splitting

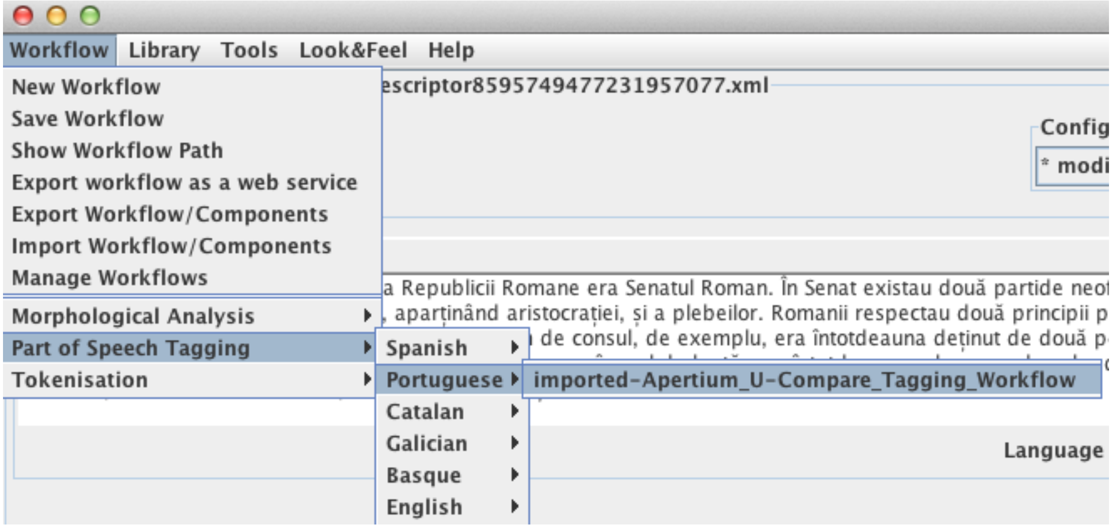
Selected Categories: Morphological Analysis, Part of Speech Tagging, Tokenisation

Category Languages:
☑ Portuguese
☑ Catalan
☑ Galician
☑ Basque
☑ English
☐ French

Figure 3: Categories/Languages box

Execution instructions

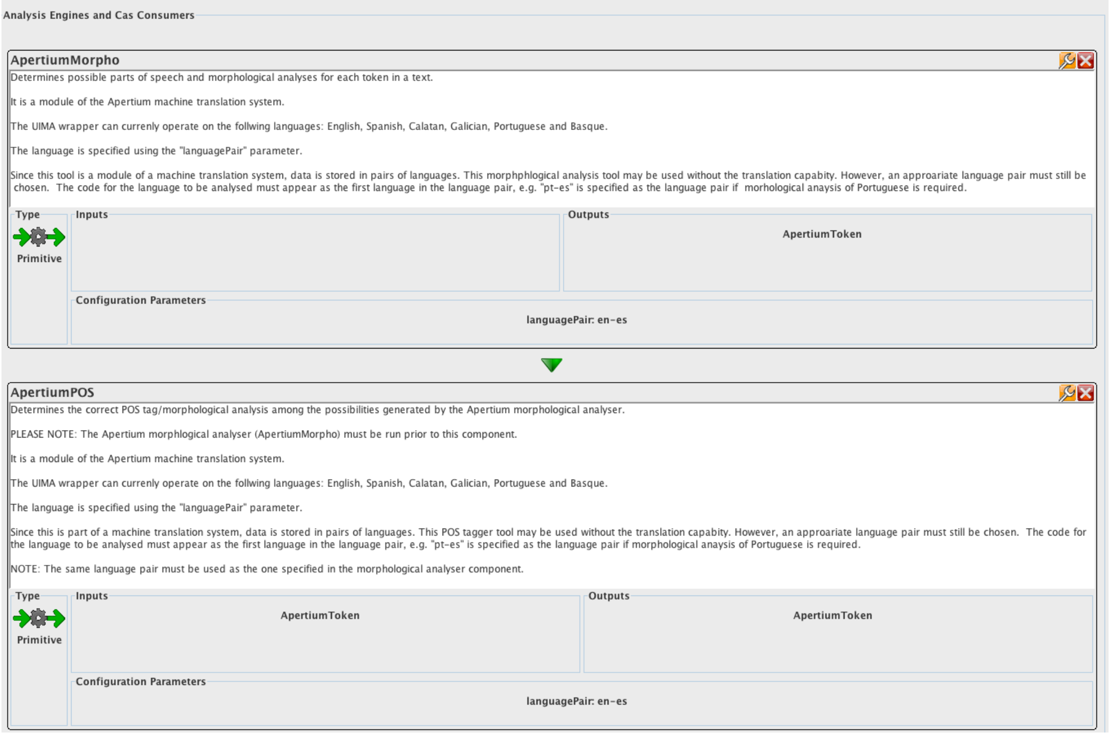
Once imported, the workflow can be loaded from the U-Compare interface. According to the steps described above, workflows are categorised according to their function and language. If a workflow supports multiple functions/languages, it will be categorised in several ways. In Figure 4, the user has decided, through the workflow menu system, that they need a workflow that can carry out part-of-speech tagging in the Portuguese language. The imported Apertium workflow is displayed as a workflow that can fulfil this requirement.



a Republicii Romane era Senatul Roman. În Senat existau două partide neo
, aparținând aristocrației, și a plebeilor. Romanii respectau două principii p
de consul, de exemplu, era întotdeauna deținut de două p

Figure 4: Selection of workflow by category and language

Selecting the workflow name will cause the workflow to be loaded into the U-Compare workflow canvas. This is shown in Figure 5.



Analysis Engines and Cas Consumers

ApertiumMorpho

Determines possible parts of speech and morphological analyses for each token in a text.

It is a module of the Apertium machine translation system.

The UIMA wrapper can currenly operate on the follwing languages: English, Spanish, Calatan, Galician, Portuguese and Basque.

The language is specified using the "languagePair" parameter.

Since this tool is a module of a machine translation system, data is stored in pairs of languages. This morphphlogical analysis tool may be used without the translation capabity. However, an approariate language pair must still be chosen. The code for the language to be analysed must appear as the first language in the language pair, e.g. "pt-es" is specified as the language pair if morhological anaysis of Portuguese is required.

Type: Primitive

Inputs:

Outputs: ApertiumToken

Configuration Parameters: languagePair: en-es

ApertiumPOS

Determines the correct POS tag/morphological analysis among the possibilities generated by the Apertium morphological analyser.

PLEASE NOTE: The Apertium morphlogical analyser (ApertiumMorpho) must be run prior to this component.

It is a module of the Apertium machine translation system.

The UIMA wrapper can currenly operate on the follwing languages: English, Spanish, Calatan, Galician, Portuguese and Basque.

The language is specified using the "languagePair" parameter.

Since this is part of a machine translation system, data is stored in pairs of languages. This POS tagger tool may be used without the translation capabity. However, an approariate language pair must still be chosen. The code for the language to be analysed must appear as the first language in the language pair, e.g. "pt-es" is specified as the language pair if morphological anaysis of Portuguese is required.

NOTE: The same language pair must be used as the one specified in the morphological analyser component.

Type: Primitive

Inputs: ApertiumToken

Outputs: ApertiumToken

Configuration Parameters: languagePair: en-es

Figure 5: Workflow loaded into the U-Compare workflow canvas

Before running the workflow, both of the components need to be configured to operate on Portuguese. This is done by clicking on the icon of both the ApertiumMorpho and ApertiumPOS components. The value of the “Language Pair” parameter should be set as “pt-es”, and then the “Confirm Changes” button should be clicked. The text for each component explains how to set the languagePair for other languages.

After having entered some text in the “Input Text Reader” component, the workflow can be run, by clicking on the Play button in U-Compare.

Input/Output data formats

Input data formats

The workflow operates on plain, unannotated text.

Output data format

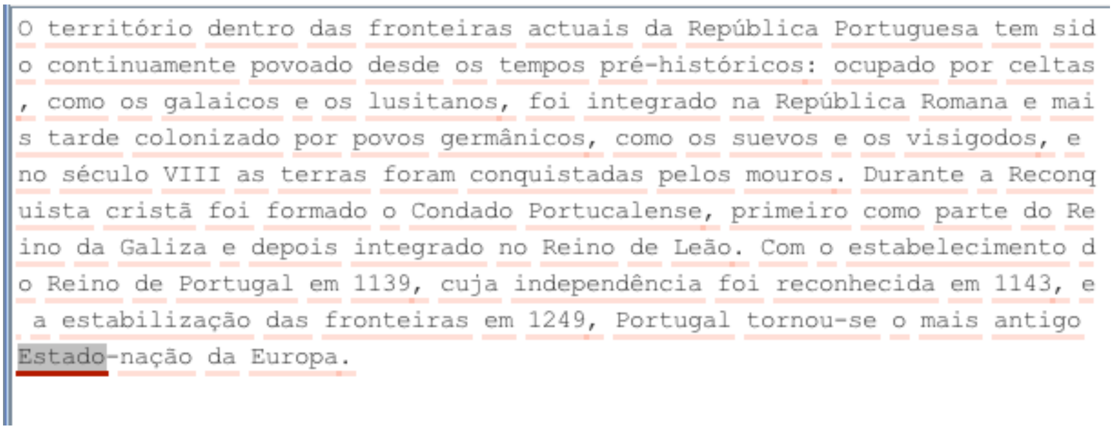
An annotation is thus added to the CAS corresponding to each token in a document, with the type "ApertiumToken", which stores the part of speech tag, base form and additional morphological information.

Integration with external tools

As mentioned above, the workflow can only be run within U-Compare framework. However, U-Compare offers an option to export workflows as web services, making them more widely useable within other code.

3. CONTENT INFORMATION

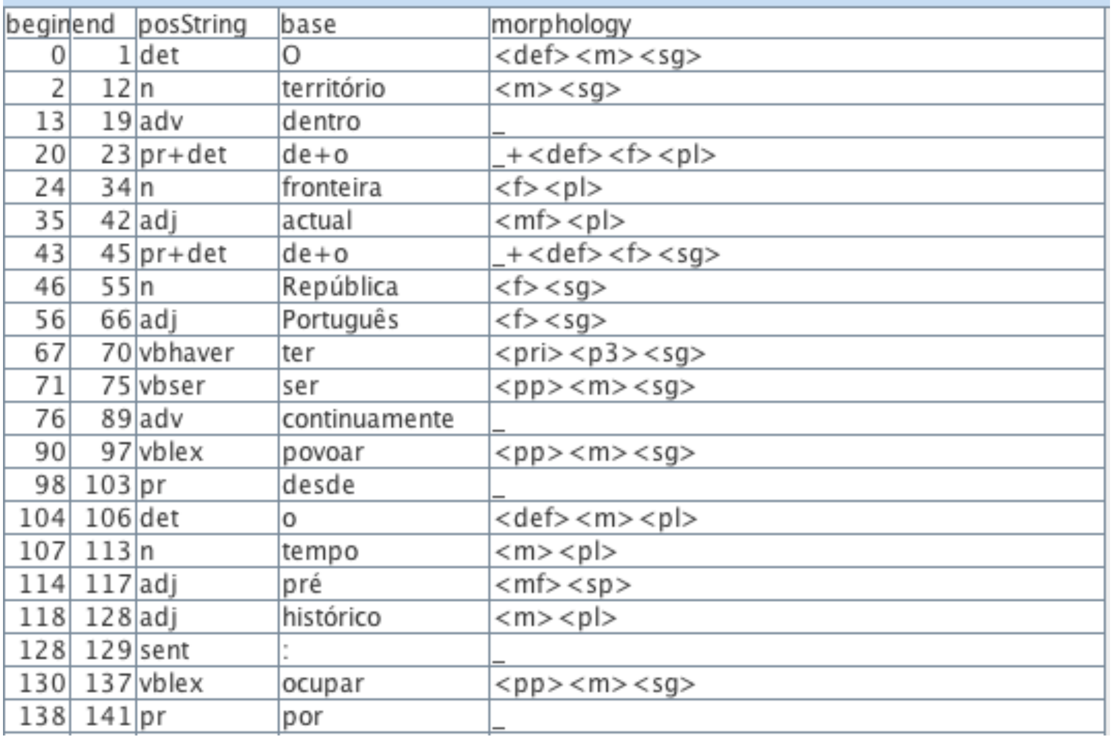
Figure 6 shows the part of the output of the workflow in the U-Compare workbench, using a sample Portuguese text. Each token in the text has been separated identified.



```
O território dentro das fronteiras actuais da República Portuguesa tem sid
o continuamente povoado desde os tempos pré-históricos: ocupado por celtas
, como os galaicos e os lusitanos, foi integrado na República Romana e mai
s tarde colonizado por povos germânicos, como os suevos e os visigodos, e
no século VIII as terras foram conquistadas pelos mouros. Durante a Reconq
uista cristã foi formado o Condado Portucalense, primeiro como parte do Re
ino da Galiza e depois integrado no Reino de Leão. Com o estabelecimento d
o Reino de Portugal em 1139, cuja independência foi reconhecida em 1143, e
 a estabilização das fronteiras em 1249, Portugal tornou-se o mais antigo
Estado-nação da Europa.
```

Figure 6: Output of the Apertium Part-of-Speech Tagging workflow

Figure 7 shows another part of analysis displayed in the U-Compare interface, with the details for each ApertiumToken annotation. These consist of the start and end offsets of each token, together with the part-of-speech tag (posString column), base form ad additional morphological information. Words that are unknown by the system are marked with a "*".



begin	end	posString	base	morphology
0	1	det	O	<def><m><sg>
2	12	n	território	<m><sg>
13	19	adv	dentro	_
20	23	pr+det	de+o	_+<def><f><pl>
24	34	n	fronteira	<f><pl>
35	42	adj	actual	<mf><pl>
43	45	pr+det	de+o	_+<def><f><sg>
46	55	n	República	<f><sg>
56	66	adj	Português	<f><sg>
67	70	vbhaver	ter	<pri><p3><sg>
71	75	vbser	ser	<pp><m><sg>
76	89	adv	continuamente	_
90	97	vblex	povoar	<pp><m><sg>
98	103	pr	desde	_
104	106	det	o	<def><m><pl>
107	113	n	tempo	<m><pl>
114	117	adj	pré	<mf><sp>
118	128	adj	histórico	<m><pl>
128	129	sent	:	_
130	137	vblex	ocupar	<pp><m><sg>
138	141	pr	por	_

Figure 7: Annotation detail table

4. LICENCES

a) The ApertiumMorpho UIMA wrapper is licensed using the GNU General Public License version 2.0 (GPLv2). Please see "COPYING.txt" in the "Licences" directory. Please acknowledge the National Centre for Text Mining, University of Manchester if you use the ApertiumMorpho UIMA component

b) The underlying Apertium software is licensed using the GNU General Public License version 2.0 (GPLv2). Please see "COPYING.txt" in the "Licences" directory.

c) The UIMA framework is licenced using the Apache licence. Please see "Apache.txt" in the "Licences" directory.

5. ADMINISTRATIVE INFORMATION

Contact

Contacts for the Apertium system can be found here:
http://wiki.apertium.org/wiki/Contact

For further information regarding this UIMA wrappers for the Apertium toools, please contact Sophia Ananiadou:

sophia.ananiadou@manchester.ac.uk

6. REFERENCES

Armentano-Oller, C., Carrasco, R., Corbí-Bellot, A.,Forcada, M., Ginestí-Rosell, M., Ortiz-Rojas, S. (2006). Open-source Portuguese–Spanish machine translation. Computational Processing of the Portuguese Language ,50-59

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.